深層ニューラルネットワークの圧縮可能性を用いた非圧縮ネットワークの汎化誤差解析

鈴木大慈: 東京大学大学院情報理工学系研究科，理研 AIP.

■**概要**　深層学習の汎化誤差理論における重要な問題として，そのパラメータ数がサンプルサイズよりも大きいにもかかわらず汎化する（過学習しない）という問題がある．VC 次元を用いた古典的な学習理論をそのまま適用すると，パラメータ数がサンプルサイズより大きい場合は汎化誤差が小さくなることが保証されない．しかし，現実の深層学習は陰的正則化等の作用で，訓練データに完全にフィットできる表現力を持ちつつも過学習を避けていることが実験的に確認されている．この状況を理論的に説明するために様々なアプローチが提案されているが，その中でも圧縮可能性を用いた汎化誤差バウンド（圧縮型バウンド）は有用なアプローチである [1, 2]．しかし，圧縮型バウンドは圧縮されたネットワークの汎化誤差を解析するもので，圧縮前の学習済みネットワークに関しては評価を与えない．本研究では，圧縮型バウンドをもとの圧縮していないネットワークに変換する一般的な枠組みを与え，複数の具体例でそのバウンドを導出する．導出したバウンドは，十分大きなサンプルサイズにおいて既存の圧縮型バウンドのバイアス項を改善する．これによって，データに依存するよりタイトな評価が得られ，データ非依存型の VC 次元などのバウンドよりもはるかにタイトな評価が得られる．具体例として，学習された重み行列や中間層の分散共分散行列がほぼ低ランクの場合に導出されたバウンドを計算し，上界を改善していることを確認する．

■**問題設定**　観測データを $D_n = (z_i)_{i=1}^n = (x_i, y_i)_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ とする．あるネットワーク $f : \mathbb{R}^d \to \mathbb{R}$ の性能を評価するために損失関数 $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ を用意し，訓練誤差および期待誤差を $\widehat{\Psi}(f) := \frac{1}{n}\sum_{i=1}^n \psi(y_i, f(x_i))$，$\Psi(f) := \mathrm{E}[\psi(Y, f(X))]$ と定義する．ただし，期待値はデータの真の分布 $(X, Y) \sim P$ に関して取る．本研究では，ある推定量 $\widehat{f}$ に対して，汎化ギャップ $\Psi(\widehat{f}) - \widehat{\Psi}(\widehat{f})$ を抑える．今，$g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ に対して，経験的な L_2-ノルムを $\|f\|_n^2 := \sum_{i=1}^n g(x_i, y_i)^2/n$ と定義し，真の分布による L_2-ノルムを $\|g\|_{L_2(P(X))}^2 := \mathrm{E}_{(X,Y)\sim P}[g(X, Y)^2]$ とする．第 ℓ 層の横幅が m_ℓ で深さが L の深層ニューラルネットモデルを $\mathcal{F}$ と書く．

■**主結果**　推定量 $\widehat{f}$ の汎化性能を調べるために，学習したネットワークの集合 $\widehat{\mathcal{F}} \subset \mathcal{F}$ とそれを圧縮したネットワークの集合 $\widehat{\mathcal{G}} \subset \mathcal{F}$ を用意する．ここで，$\widehat{f}$ を「圧縮する」ことでより単純な $\widehat{g} \in \widehat{\mathcal{G}}$ を得るというのは，パラメータを削減するなどして，より小さなネットワーク $\widehat{g}$ を生成することとする．$\widehat{\mathcal{F}} - \widehat{\mathcal{G}} := \{f - g \mid f \in \widehat{\mathcal{F}}, g \in \widehat{\mathcal{G}}\}$ として，$\widehat{\mathcal{F}} - \widehat{\mathcal{G}}$ の局所 Rademacher 複雑度を $\dot{R}_r(\widehat{\mathcal{F}} - \widehat{\mathcal{G}}) := \bar{R}_n(\{h \in \widehat{\mathcal{F}} - \widehat{\mathcal{G}} \mid \|h\|_{L_2(P(X))} \le r\})$ と定義し (ただし，$\bar{R}_n$ は関数集合の Rademacher 複雑度とする)，これに対してある関数 $\phi : [0, \infty) \to [0, \infty)$ が存在して以下が成り立つと仮定する: $\dot{R}_r(\widehat{\mathcal{F}} - \widehat{\mathcal{G}}) \le \phi(r)$ かつ $\phi(2r) \le 2\phi(r)$ $(\forall r > 0)$. さらに，$t > 0$ に対して $r_*(t) := \inf\left\{r > 0 \mid 8\frac{\phi(r)}{r^2} + M\sqrt{\frac{4t}{r^2 n}} + M^2\frac{2t}{r^2 n} \le \frac{1}{2}\right\}$ と定義する．すると次の定理を得る．

Theorem 1. $\widehat{f}$ と $\widehat{g}$ の経験的な L_2-距離がある $\hat{r} > 0$ を用いて $\|\widehat{f} - \widehat{g}\|_n \le \hat{r}^2$ $(a.s.)$ のように抑えられているとする．また，$t \ge 1$ に対して $\dot{r} := \sqrt{2(\hat{r}^2 + r_*(t)^2)}$ とすると，ある適当な仮定のもと，以下の不等式が確率 $1 - 3e^{-t}$ で成り立つ:

$$\Psi(\widehat{f}) \le \widehat{\Psi}(\widehat{f}) + C\left[\bar{R}_n(\widehat{\mathcal{G}}) + \sqrt{\tfrac{t}{n}} + \dot{R}_{\dot{r}}(\widehat{\mathcal{F}} - \widehat{\mathcal{G}})\log(n)^{\frac{3}{2}} + \dot{r}\sqrt{\tfrac{t}{n}} + \tfrac{1+t}{n}\right].$$

特に，学習された重み行列 $W^{(\ell)}$ と中間層からの出力の分散共分散行列がほぼ低ランクである場合（それぞれの第 j 番目に大きな特異値が $Cj^{-\alpha}$ および $Cj^{-\beta}$ で抑えられる場合），汎化誤差は $\Psi(\widehat{f}) - \widehat{\Psi}(\widehat{f}) \lesssim \sqrt{\frac{(\sum_{\ell=1}^L m_\ell)^{\frac{4/\beta}{4/\beta+2(1-1/2\alpha)}}}{n}} + \left(\frac{\sum_{\ell=1}^L m_\ell}{n}\right)^{\frac{2\alpha}{1+2\alpha}}$ といった上界で抑えられる．

References

[1] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang. Stronger generalization bounds for deep nets via a compression approach. ICML2018, pp. 254–263, 2018.

[2] T. Suzuki, H. Abe, T. Murata, S. Horiuchi, K. Ito, T. Wachi, S. Hirai, M. Yukishima, and T. Nishimura. Spectral-Pruning: Compressing deep neural network via spectral analysis. arXiv:1808.08558, 2018.